CongloMed

Conglomerate Medical
Data Warehousing for Analysis, Mining, and Classification

Colby Ford, 2015

The University of North Carolina at Charlotte

The Onslaught of Big Data in Medicine

With “Big Data” being the coined term that industries from finance to manufacturing to retail are using to describe the massive pile of information being collected about clients, transactions, etc., it is obvious that there is a great demand for data integration such that understanding of the data can take place.

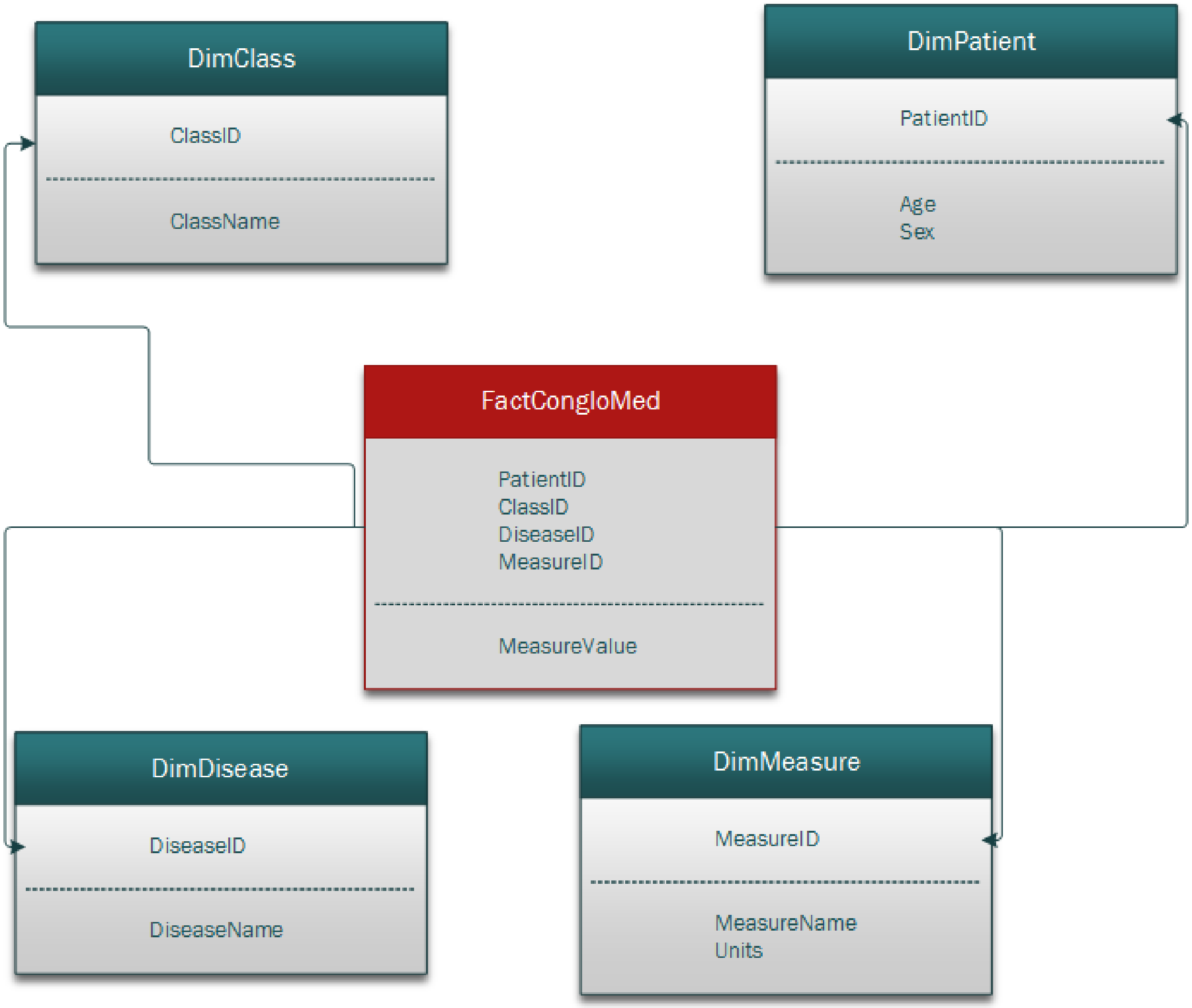
The medical field is no different. Patients come into medical facilities and medical data is stored about them every step of the way. Blood pressure readings, x-rays, diagnoses, are all charted in a medical facility’s database. In this regard, it is far from living up to its potential.

If medical facilities could join their immense amounts of data together, better research could take place. No longer would data be in silos from one hospital to the other and even from one department to the other. How interesting would it be if we could correlate test results from a nephrologist to data from a diabetes specialist and see the overlaps in trends? Maybe we could dive into the correlation and form research around the cause and effect relationship with high blood glucose and renal failure. The possibilities are limitless if the data could be available.

How Conglomerate Medical (ConggloMed) Data Warehousing Works

Data are placed into the data warehouse in a fashion such that all diseases, patients, measures (readings from tests), etc. are all in one place. This way, cross-sectional analysis can occur between patients (patients from different hospitals, for example), diseases (lung cancer vs. breast cancer), and trends of all diseases by patients' age or sex.

For the example here, the warehouse schema is as follows:



Class is the dimension where the outcome would be housed. Think of this as the diagnosis. Example entries here would include, positive, negative, benign, malignant, present, absent, etc.

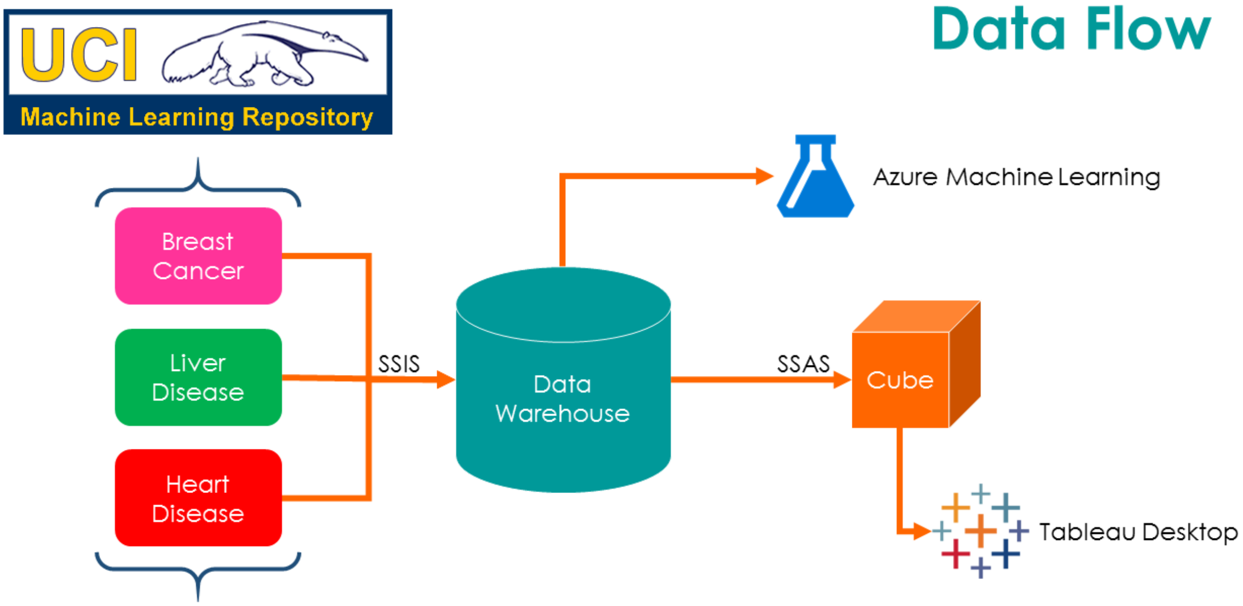
The Disease dimension can be expanded to include disease categories, subtypes, etc. as well as descriptors and more.

The Measure dimension houses all possible tests any patient has. For example, this would include blood pressure, blood glucose, presence of a particular antibody, etc.

The Patient dimension will house demographic information about each patient such as age, sex, race, income, etc.

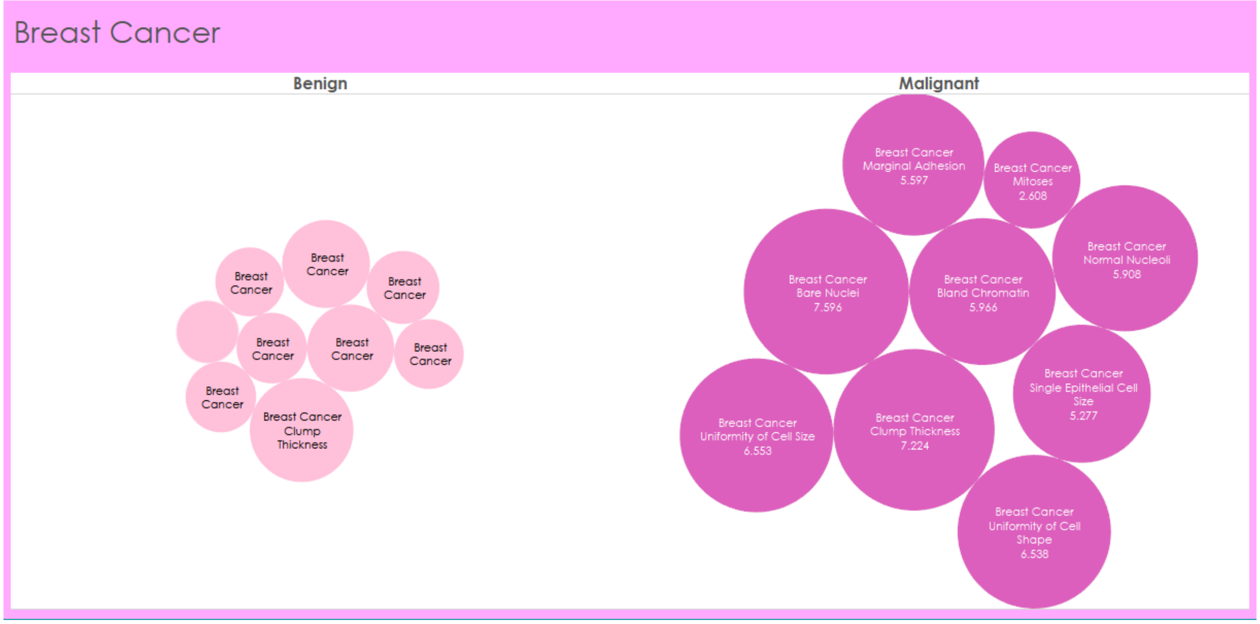
Analysis and Classification

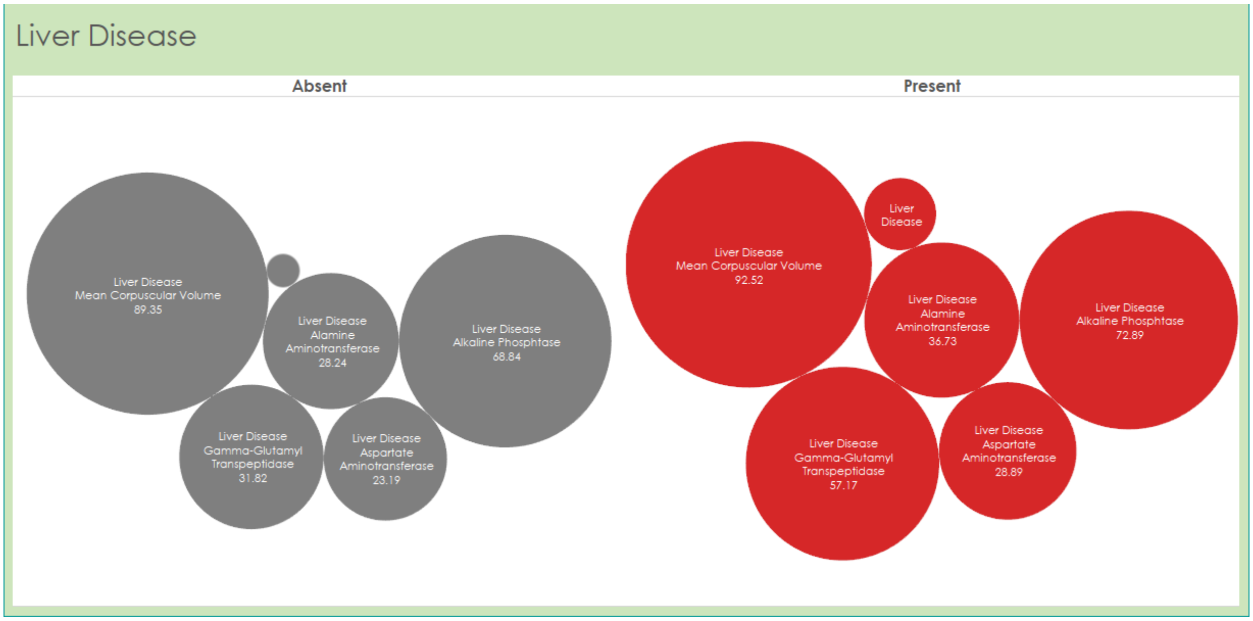
By housing data in a data warehouse, we can create an OLAP cube for quick analytical processing. Plus, querying a cube or otherwise connecting to it via third party software is very simple.

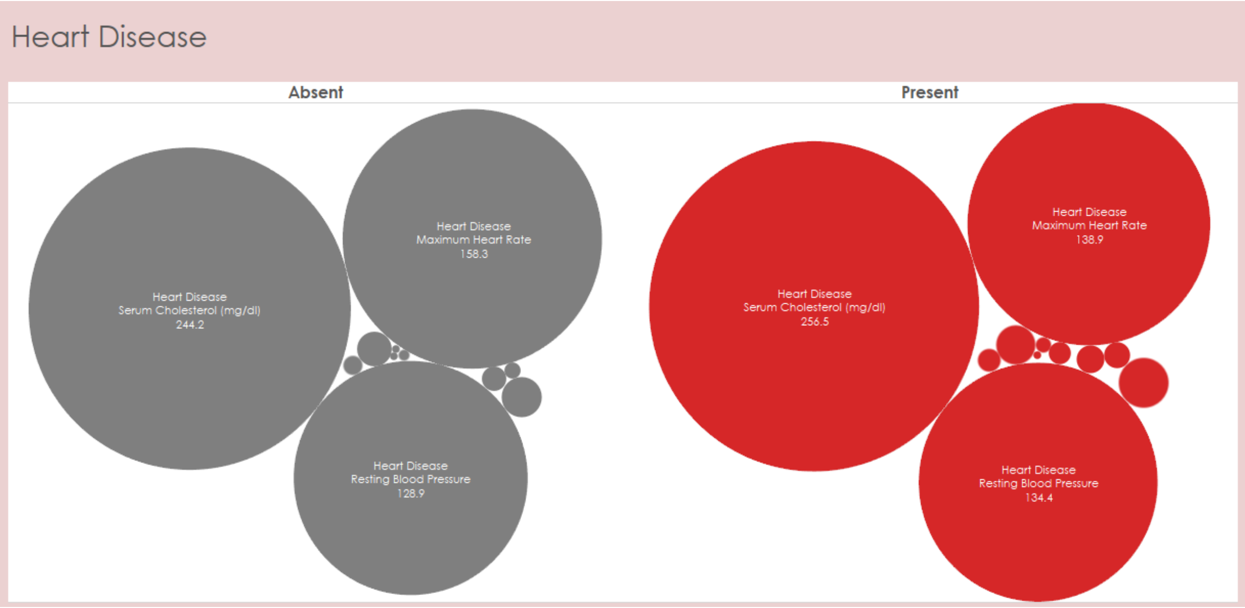


The sample datasets come from the UCI Machine Learning Repository. Breast Cancer, Liver Disease, and Heart Disease are loading into the cube and can then be used in analysis or visualization software. Here, I use Tableau Desktop and Microsoft Azure Machine Learning.

Visualization





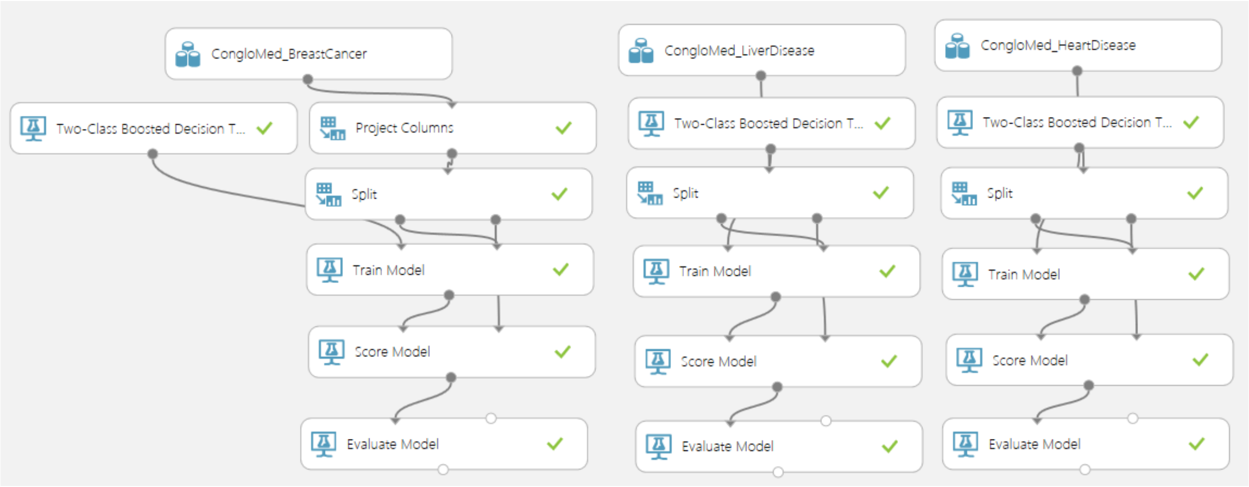


By using Tableau, we can view the average measure values as they related to each disease and then also split them by whether or not the individual has the disease. Here, we can see that the measure related to breast cancer can really show a heavy correlation with malignancy whereas the measure values in heart disease and liver disease do not change much if the patient actually has the disease versus patients who do not.

Machine Learning

The sample datasets, which came from the UCI Data Repository are for use in classification examples. In this case, whether or not the patient has the disease given the other measures.

Here, we can use a solution like Microsoft Azure Machine Learning to apply a two-class boosted decision tree algorithm to the data (split by each disease).



Once these models are built, future data can be passed through the models using the API or web service that Azure Machine Learning generates. Without a conglomerate data source, the analysis would have to take place in silos (one for breast cancer, one for liver disease, etc.) whereas here, we can run the analysis much faster and in parallel.

Conclusion

Medical research is an import part in developing new medicines, cures, and understanding about the afflictions that humans face. Diseases, genetics, and treatments are all very diverse and often hard to keep track of. By using a solution like CongloMed, data mining and research can occur at an unhindered rate. With all data about many diseases, patients, treatments, all in one place, there's a much greater opportunity for scientific breakthrough!